Rekonstruktion von abgeleiteten Variablen mittels zeilen- bzw. satzübergreifender Operationen in STATA im Mikrozensus

John, Kristina

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
John, K. (2007). *Rekonstruktion von abgeleiteten Variablen mittels zeilen- bzw. satzübergreifender Operationen in STATA im Mikrozensus.* (GESIS-Methodenberichte, 3/2007). Mannheim: GESIS-ZUMA. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-206828

GESIS-ZUMA-Methodenbericht Nr. 3/2007

Rekonstruktion von abgeleiteten Variablen mittels zeilen- bzw. satzübergreifender Operationen in STATA im Mikrozensus

Kristina John

Dezember 2007

ISSN 1865-7575

GESIS-ZUMA
German Microdata Lab (GML)
Quadrat B2,1
Postfach 12 21 55
68072 Mannheim
Telefon: (0621) 1246-252
Telefax: (0621) 1246-100
E-Mail: Kristina.John@gesis.org

GESIS-Methodenberichte

Die GESIS ist ein Institut der Leibniz-Gemeinschaft.

ISSN: 1865-7575

Herausgeber, Druck GESIS

und Vertrieb: Postfach 12 21 55, 68072 Mannheim

Printed in Germany

Inhaltsverzeichnis:

Zusammenfassung

Der Mikrozensus eignet sich als Datengrundlage für sowohl arbeitsmarkts- und bevölkerungsspezifische als auch familiensoziologische Fragestellungen. Allerdings werden lediglich acht der 150 haushalts- und familienrelevanten Merkmale direkt erhoben. Die für viele demographische Analysen unabdingbaren Kontextinformationen über eine Familie oder einen Haushalt, wie beispielsweise die Anzahl der Kinder unter drei Jahren im selbigen, werden erst im nachhinein von den statistischen Ämtern generiert.

Ziel des vorliegenden Berichts ist es, beispielhaft aufzuzeigen, wie derartige *abgeleitete Variablen* selbst generiert werden können. Das Erlernen der eigens durchgeführten Rekonstruktion von Aggregatinformationen über einen Haushalt oder eine Familie bietet der Forscherin oder dem Forscher den Vorteil, der eigenen wissenschaftlichen Fragestellung entsprechend selbst Kontextinformationen über die Untersuchungseinheit zu generieren und eigene themenspezifische Filter zur Abgrenzung der interessierenden Variablen zu wählen. Im Folgenden wird anhand der beiden Variablen „*EF524* Erwerbslose (EU-Definition) im Haushalt“ und „*EF557* Geschlecht: Haushaltsbezugsperson“ gezeigt, wie *abgeleitete Variablen* mithilfe von STATA selbst erzeugt werden können. Dabei bilden die drei Befehle „*egen*“, „*merge*“ und „*joinby*“ unterschiedliche Optionen der Umsetzung. Ihre Vor- und Nachteile werden kritisch erörtert und anschließend dem Pendant in SPSS gegenüber gestellt.

1 Einleitung

Der Mikrozensus erhebt jährlich 1% der Privathaushalte in Deutschland ist sowohl eine rotierende Panelstichprobe als auch eine Flächen- bzw. Klumpenstichprobe, in der die Haushalte eines Auswahlbezirks vier Jahre lang befragt werden. Jedes Jahr wird ein Viertel der Auswahlbezirke (Klumpen) ausgetauscht. Aufgrund, des resultierenden großen Stichprobenumfangs und der fast vollständigen Ausschöpfung bieten Mikrozensusdaten vielfältige Auswertungsmöglichkeiten:

Neben Informationen zu Bevölkerung und Arbeitsmarkt stehen ForscherInnen Angaben für familienwissenschaftliche Zwecke zur Verfügung. Jedoch werden nur wenige dieser haushalts- und familienbezogenen Variablen direkt erhoben. Erst durch so genannte *abgeleitete Variablen,* die post-hoc vom Statistischen Bundesamt durch die Kombination der im Mikrozensus direkt erhobenen Merkmale generiert werden, entwickelt sich das familiensoziologische Potential der Daten.

In diesem Papier sollen verschiedene Möglichkeiten der Rekonstruktion so genannter Bandsatzerweiterungen mittels zeilenübergreifender Operationen in STATA[1] und ihre „Übersetzung“ in SPSS illustriert werden. Die verschiedenen Optionen der Syntaxgestaltung werden der Einfachheit halber anhand der Replikation der bereits im Scientific Use File des Mikrozensus 2004 vorhandenen Bandsatzerweiterungen „*EF524* Erwerbslose (EU-Definition) im Haushalt“ und „*EF557* Geschlecht: Haushaltsbezugsperson“ verdeutlicht. Auf diese Weise kann die prinzipielle Vorgehensweise erläutert werden, die sich dann auf andere Anwendungen übertragen lässt.[2]

Der Datensatz, der diesen Bandsatzerweiterungen zugrunde liegt, ist das Mikrozensus Scientific Use File (SUF) 2004.[3] Als Haushaltsstichprobe enthält er eine Vielzahl an Informationen auf der Ebene von Haushalten, Familien und Lebensgemeinschaften, wie z. B. die Zahl der Kinder in verschiedenen Altersgruppen in der Familie oder im Haushalt der oder Beruf der Bezugsperson des Haushalts. Für sozialwissenschaftliche Analysen sind solche Merkmale eine wichtige Ergänzung zu den Personenangaben, da sie soziale und wirtschaftliche Kontexte des individuellen Handelns widerspiegeln. In den SUF des Mikrozensus liegen bereits viele von den statistischen Ämter routinemäßig erzeugte abgeleitete Variablen ("derived variables") oder so genannte Bandsatzerweiterungen sowie Typisierungen auf Haushalts- oder Familienebene vor. In einem Methodenbericht haben Lengerer und Boehle (2006) u.a. auch anhand der Replikation der Variable *EF524* gezeigt, wie mit SPSS solche Variablen gebildet werden können, falls sie nicht bereits im Scientific Use File vorliegen oder je nach Fragestellung eine andere Abgrenzung oder Rekodierung als die der statistischen Ämter präferiert wird.[4]

Ergänzend dazu und darauf aufbauend wird in diesem Papier gezeigt, wie diese abgeleiteten Variablen mit STATA konstruiert werden können: Es wird dargestellt, wie auf Haushaltsebene, also "satzübergreifend" Informationen aggregiert und dieses Ergebnis den einzelnen Sätzen, d.h. Personen im Haushalt, wieder zugespielt wird. Die hierfür mit STATA verwendbaren Kommandos sind *"egen"*, *"merge"* und *"joinby"*, deren ausführliche Syntaxabfolge im Anhang als auch mithilfe der auf der GML-Homepage befindlichen Syntaxdatei des Programmes SPSS nachvollzogen werden kann.

---

[1] Die folgenden Ausführungen sind auf die STATA-Versionen 9 und 10 anwendbar.

[2] An dieser Stelle möchte ich mich bei denen herzlich bedanken, die diese Arbeit durch Anregungen unterstützt haben: Andrea Lengerer, Bernhard Schimpl-Neimanns, Julia Schroedter und Heike Wirth.

[3] Siehe die Datenbeschreibung unter http://www.gesis.org/Dauerbeobachtung/ GML/Daten /MZ/mz_2004/index.htm).

[4] Siehe unter http://www.gesis.org/Dauerbeobachtung/GML/Service/Mikrodaten-Tools/Bandsatz96_04/index.htm

2 Das Erzeugen von Bandsatzerweiterungen mithilfe satzübergreifender Operationen in STATA

2.1 Überlegungen vor der Analyse

Vor Beginn der eigenen satzübergreifenden Arbeiten sollte, gleich welcher Befehl gewählt wird, überlegt werden, auf welcher Ebene (Haushalts-, Familien- oder Lebensgemeinschaftsebene) Kontextinformationen benötigt werden.[5] Zudem besteht die Option, entweder Informationen über eine ganze Einheit (z.B. einen Haushalt) den Personen einer Einheit oder Informationen über einer sich in der Einheit befindenden Person den anderen Personen dieser Einheit zuzuspielen.[6] Das Zuspielen der Informationen zu allen dieser Einheit zugehörigen Individuen soll im Folgenden anhand der Rekonstruktion der Variable „*EF524* Erwerbslose (EU-Definition) im Haushalt" illustriert werden, während das Zuspielen der Information über eine bestimmte Person dieser Einheit zu allen Mitgliedern dieser Einheit anhand der Rekonstruktion der Variable „*EF557* Geschlecht: Haushaltsbezugsperson" verdeutlicht werden soll. Spielt man die Information über die Anzahl der Erwerbslosen im Haushalt, ergo eine Information über den Haushalt, den einzelnen Haushaltsmitgliedern zu, ergibt sich für einen vierköpfigen Haushalt beispielsweise Folgendes:

Abb. 1: Vergleich von *V524* und *EF524* am Beispiel eines Vier-Personen-Haushaltes

EF1 Bundesland	***EF3*** Auswahlbez.	***EF4*** Haushalts-Nr.	***EF504*** Erwerbsstatus	***EF506*** Bev.: Privathaushalt	***EF524*** N Erw.lose	***V524*** N Erw.lose
1	369	4	2 [erwerbslos]	1	2	2
1	369	4	1 [erwerbstätig]	1	2	2
1	369	4	4 [Nichterwerbsperson]	1	2	2
1	369	4	2 [erwerbslos]	1	2	2

Alle vier Personen weisen dieselbe Merkmalskombination der Variablen *EF1*, *EF3* und *EF4* auf, da sie in einem Haushalt leben. Die erste und die letzte Person sind erwerbslos (*EF504==2*), während die zweite Person erwerbstätig ist (*EF504==1*) und die dritte Person als Nichterwerbsperson (*EF504==4*) definiert wird. Da sich unsere Analyse auf die Haushaltsebene bezieht, wird nur die Bevölkerung in Privathaushalten betrachtet (*EF506==1*). Es handelt sich also um zwei erwerbslose Personen in diesem Haushalt. Diese Information wird, wie in dieser Tabelle zu sehen, allen

[5] Je nachdem, welche Ebene benötigt wird, greift ein anderes Bevölkerungskonzept. Beispielsweise wird für Auswertungen auf der Ebene des Haushaltes eine Einschränkung auf die Bevölkerung in Privathaushalten (*EF506==1*) getroffen.

[6] So kann man z.B. den Beruf der Bezugsperson im Haushalt allen anderen Haushaltsmitgliedern zuspielen

Mitgliedern des Haushaltes zugespielt. Da die Bandsatzerweiterung *v524* korrekt gebildet wurde, stimmt sie mit der vom Statistischen Bundesamt generierten Variable *EF524* überein.
Wird hingegen die Information über das Geschlecht der Haushaltsbezugsperson den einzelnen Mitgliedern des Haushaltes zugespielt, erhält jedes Mitglied dieses Vier-Personen-Haushaltes den Wert 1 [vgl. Abb. 2]. Dies bedeutet, dass jeder Haushaltsperson die Information zugewiesen wird, dass die Haushaltsbezugsperson ein Mann ist.

Abb. 2: Vergleich von *V557* und *EF557* am Beispiel eines Vier-Personen-Haushaltes

EF1 Bundesland	***EF3*** Auswahlbez.	***EF4*** Haushalts -Nr.	***EF32*** Geschlecht	***EF506*** Bev.: Privathaus-halt	***EF507*** Haushaltsbezugsperson	***V557*** Geschl:HH-Bezp.	***EF557*** Geschl: HH-Bezp.
1	369	4	1[Männlich]	1	1 [Bezugspersohn]	1 [Männlich]	1
1	369	4	2 [Weiblich]	1	2 [Ehegattin]	1 [Männlich]	1
1	369	4	2 [Weiblich]	1	3 [Tochter]	1 [Männlich]	1
1	369	4	2 [Weiblich]	1	3 [Tochter]	1 [Männlich]	1

Im Folgenden wird erläutert, wie man die abgeleiteten Variablen *EF524* und *EF557* mit STATA-Kommandos replizieren kann.

2.1.1 Der Befehl „egen“ in STATA

Der Befehl „*extended generate*“ (abgekürzt „*egen*“) in STATA bietet die eleganteste Lösung, um an Kontextinformationen zu gelangen, da er die Aggregation und Zuweisung in einem Schritt erledigt und keine zuvorige Sortierung des Datensatzes benötigt.
Z.B. für die Anzahl der erwerbslosen Personen in einem Haushalt:

```
egen [type] newvar = fcn(arguments) [if] [in] [, options]
```

```
egen v524 = total(EF504==2 & EF506==1), by (EF1 EF3 EF4)
```
[7]

oder für das Geschlecht der Haushaltsbezugsperson, wobei zunächst sichergestellt wird, dass nur Fälle mit einer Haushaltsbezugsperson berücksichtigt werden:[8]

```
egen n_bezper_hh = total(EF506==1 & EF507==1), by(EF1 EF3 EF4)
```

```
if n_bezper_hh==1 {
```

[7] EF504: Erwerbstyp (2 = Erwerbslose, sofort verfügbar (EU-Definition)); EF506: Bevölkerung: Privathaushalte (1 = Bev. in Privat-HH = Personen, die zur Bevölkerung in Privathaushalten gehören).

[8] In einzelnen Fällen treten Haushalte mit mehreren Bezugspersonen auf, wobei es sich um Datenfehler handelt.

```
egen v557 = max(EF32 * (EF506==1 & EF507==1) ), by(EF1 EF3 EF4)
}
else {
replace v557==.
}
label variable v557 "Geschlecht: Haushaltsbezugsperson"
```

Die Befehlsstruktur von *„egen“* gleicht der von *„generate“* (*„gen“*). Jedoch beinhaltet sie eine große und ständig wachsende Zahl von Erweiterungen, d.h. eigentlich eine Reihe von hintereinander ausgeführten, mehr oder weniger komplizierten *„generate“*- und *„replace“*-Kommandos und lässt sich für diverse Operationen verwenden (vgl. Kohler und Kreuter 2006: 94).

Nach dem Befehl folgt der Name der Variable (hier: *v524* oder *v557*), die erzeugt werden soll, dann ein Gleichheitszeichen und schließlich eine *„egen“*-Funktion. Bei der Rekonstruktion der Variable *EF524* beispielsweise zeichnet sich die Funktion dadurch aus, dass innerhalb der Klammer die für unsere Frage benötigte Abgrenzung auf Individualebene, also die Einschränkung auf die Bevölkerung in Privathaushalten (*EF506*==1), erfolgt. Die Ausprägung *EF506* ungleich 1 würde bedeuten, dass die Funktion sich auf die Bevölkerung in Gemeischaftsunterkünften bezieht. Die der Klammer vorstehende Funktion *„total“* bewirkt, dass der Gesamtwert der erwerbslosen Personen im Haushalt gebildet wird.[9] Da die Ausprägungen der Beobachtungen hinsichtlich der Betroffenheit von Erwerbslosigkeit (*EF504* == ==2) binär mit 1 und 0 codiert werden, kann die Summe der Erwerbslosen in einem Haushalt jeweils korrekt durch die Kombination der Variablen *EF1, EF3* und *EF4* gebildet werden.

Mithilfe des *„list“*-Befehls kann man jetzt beispielsweise feststellen, ob die erzeugte Bandsatzerweiterung *v524* zu den Beobachtungen, sortiert nach den Zuweisungsvariablen *EF1, EF3* und *EF4*, kongruent ist:

```
list [varlist] [if] [in] [, options]

list EF1 EF3 EF4 EF504 EF506 EF524 v524, nolab sepby(EF1 EF3 EF4)
```

Da diese Liste aufgrund der hohen Fallzahl im Mikrozensus sehr lang und unpraktisch wird, überprüft man, ob die Übereinstimmung der beiden Variablen für alle Haushalte gilt. Erkennbar ist

[9] Bei STATA Version 8 wird anstelle des Befehls „total“ der Ausdruck „sum“ verwendet.

dies beispielsweise anhand der Diagonale, die sich ergibt, wenn man eine Kreuztabelle aus *EF524* und *v524* erzeugt:

```
tab v524 EF524 if EF506==1, miss
```

Abb. 3: Kreuztabelle der Variablen *EF524* und *V524*

Anzahl ILO-Erwerbslose im Privathaushalt *(V524)*	Erwerbslose (EU-Definition) im Haushalt: Anzahl *(EF524)* 0	1	2	3	4 o. mehr	Insgesamt
0	436.859	0	0	0	0	436.859
1	0	50.371	0	0	0	50.371
2	0	0	6.917	0	0	6.917
3	0	0	0	608	0	608
4 oder mehr	0	0	0	0	49	49
Total	436.859	50.371	6.917	608	49	494.804

Entsprechend ergibt sich folgende Kreuztabelle für unser anderes Beispiel:

Abb. 4: Kreuztabelle der Variablen *EF524* und *V524*

Geschlecht: Haushalts-bezugsperson (*V557*)	Geschlecht: Haushaltsbezugsperson (*EF557*) männlich	weiblich	Insgesamt
1	380.118	0	
2	0	114.686	
Total	380.118	114.686	494.804

2.1.2Der Befehl „merge“ in STATA

Eine weitere Möglichkeit, Bandsatzerweiterungen zu erzeugen, bietet der Befehl *„merge“*. Charakteristisch für das *„Merge“*- Kommando ist, dass es zwei Datensätze, von denen ein Datensatz aggregierte Information und der andere die ursprünglichen Individualdaten der Beobachtungen enthält, zusammenspielt. Die Zuweisung der interessierenden Information verläuft jeweils über mindestens eine, in unserem Beispiel drei (*EF1*, *EF3* und *EF4*) Schlüsselvariablen, nach denen die Daten sortiert werden und daher in einer fixen Reihenfolge untereinander stehen. Nur durch diese Ordnung kann die Zuweisung korrekt erfolgen, weil die Kenntnis, welche Personen miteinander in einem Haushalt, in einer Familie oder in einer Lebensgemeinschaft leben eine wesentliche Voraussetzung für die Konstruktion von abgeleiteten Variablen ist. Da sich der

Identifikator Haushaltsnummer (*EF4*) auf den jeweiligen, hierarchisch übergeordneten Auswahlbezirk (*EF3*) bezieht, müssen die Beobachtungen entsprechend sortiert werden. In unserem Beispiel möchten wir untersuchen, wie hoch jeweils die Anzahl der erwerbslosen Personen in einem Haushalt ist. Folglich ist die zu konstruierende Bandsatzerweiterung auf der Haushaltsebene angesiedelt. Für die korrekte Zuweisung der Beobachtungen zu ihren Haushalten sind erstere nach den Variablen *EF1* „Bundesland“[10], *EF3* „Auswahlbezirk“ und *EF4* „Haushaltsnummer“ zu sortieren. Dabei müssen die drei Variablen in exakt dieser Reihenfolge aufgeführt werden, da sie eine hierarchische Struktur aufweisen: Das bedeutet, dass zunächst alle Fälle nach den verschiedenen Bundesländern sortiert werden, in den jeweiligen Bundesländern erhalten sie jeweils eine Ausprägung für den Auswahlbezirk in diesem Bundesland und in dem entsprechenden Auswahlbezirk je eine Ausprägung für die Haushaltsnummer. Dieser hierarchische Aufbau ermöglicht es, dass die Fälle anhand der Kombination der Ausprägungen dieser drei Variablen eindeutig zu den Haushalten, in denen sie leben, zugewiesen werden können. Bei der Verwendung des *„merge-“* oder des *„joinby-“* Befehls muss der Datensatz also zuvor durch einen *„sort-“* Befehl sortiert werden.[11]

„*Merge*“ bewirkt, dass einem Datensatz ein zweiter seitlich angefügt wird. Die prinzipielle Vorgehensweise ist in den Abbildungen 3 und 4 anhand des Beispiels *V542* veranschaulicht.

„*Merge*“ unterscheidet sich weiterhin dadurch von „*egen*“, dass mehr Zwischenschritte benötigt werden, um dasselbe Ergebnis zu erzielen.

Doch nun zur Syntax: Nach der Sortierung nach den Schlüsselvariablen muss das Ergebnis im Unterschied zum *„egen“*-Befehl gespeichert werden. Danach bieten sich zwei Optionen zur Konstruktion von Bandsatzerweiterungen:

Werden die Individualdaten den Aggregatdaten zugespielt, wird zunächst der ursprüngliche Datensatz nach den Zuweisungsvariablen sortiert und gespeichert (*filename1*)[12]. Es folgt die Einschränkung auf Privathaushalte, d.h. die Abgrenzung auf Individualebene. Die Aggregation der Beobachtungen auf Haushaltsebene erfolgt mithilfe des *„collapse“*-Befehls:

---

[10] Die korrekte Zuweisung würde in unserem Beispiel, d.h. unter Verwendung des Mikrozensus 2004, auch ohne die Variable *EF1* erfolgen. Da die geschilderte Vorgehensweise zur Erzeugung von Bandsatzerweiterungen jedoch auch auf der Datengrundlage anderer Mikrozensen möglich ist, wird der Ausführlichkeit halber die Variable *EF1* mit einbezogen. In den SUF der Mikrozensen 1989 bis 1997 ist jedoch die Haushaltsnummer fortlaufend, d.h. dass dort die Haushaltsnummer als Identifikator ausreichen würde.

[11] Zur Konstruktion einer Bandsatzerweiterung ist der Befehl „*merge*“ dem Befehl „*joinby*“ vorzuziehen, da letzterer umfassender und schwieriger zu handhaben ist und daher nur in Ausnahmefällen benötigt wird. Deshalb sei hier auf die Hilfefunktion und den Do-File verwiesen (vgl. Kohler und Kreuter 2001:327).

[12] Siehe Anhang S. 13.unter 2a)

```
collapse clist [if] [in] [weight] [, options]
collapse (sum) v524, by(EF1 EF3 EF4)
```

Abb. 3: Zuspielen von Aggregatdaten zu Individualdaten

EF1 **Bundesland**	***EF3*** **Auswahlbezirk**	***EF4*** **Haushaltsnummer**	***V524*** **Anzahl Erwerbslose im HH**
1	1	1	2
1	1	1	2
1	1	1	2
1	1	1	2
1	1	2	1
1	1	2	1
.			
.			
.			
.			

EF1	***EF3***	***EF4***	***V524***	**Einheit**
1	1	1	2	Haushalt 1
1	1	2	1	Haushalt 2
...	...	...	...	Haushalt i

Die entstandenen aggregierten Daten müssen erneut nach den ursprünglichen drei Zuweisungsvariablen *EF1, EF3* und *EF4* sortiert und gespeichert werden (*filename2*). Dies entspricht (zunächst) dem so genannten Master-File. Dieses enthält Beobachtungen des Datensatzes, die gegenwärtig im Arbeitsspeicher sind, d.h. alle (eventuell modifizierten) Variablen, die gegenwärtig vorhanden sind, solange kein neuer Datensatz geladen wird. Es ist vom sogenannten Using-File abzugrenzen, das die Beobachtungen des Datensatzes, die als Datei in Form eines Stata-Format-Datensatzes gespeichert sind, beinhaltet.[13] Folglich ist immer das File, das sich gerade im Arbeitsspeicher befindet, das Master-File und das übrige das Using-File.
Daraufhin wird das Individualdaten-File (*filename1*) als neues Master-File eingelesen:

[13] In diesem Fall ist es das Mikrozensus-File mz2004.dta (vgl. Anhang, S. 12ff.)

```
use filename1, clear
```

Abb. 4: Zuspielen von Individualdaten zu Aggregatdaten

Individualdaten *(filename 1)*

EF1 **Bundesland**	***EF3*** **Auswahlbezirk**	***EF4*** **Haushaltsnummer**	***V524*** **Anzahl Erwerbslose im HH**
1	1	1	2
1	1	1	2
1	1	1	2
1	1	1	2
1	1	2	1
1	1	2	1
.			
.			
.			
.			

Aggregatdaten *(filename 2)*

EF1	***EF3***	***EF4***	***V524***	**Einheit**
1	1	1	2	Haushalt 1
1	1	2	1	Haushalt 2
...	...	...	...	Haushalt i

Der anschließende merge-Befehl führt dieses neue Masterfile mit den Aggregatdaten "*filename2*" zusammen:

```
merge EF1 EF3 EF4 using filename2, uniqusing
```

Im Unterschied zu dieser Option ist die Zuspielung von Aggregatdaten zu Individuladaten unkomplizierter, da ein Kommando weniger benötigt wird (Vgl. Abb. 5, S. 10). Nach dem *„collapse“*- Befehl entfällt die erneute Sortierung und Speicherung. Der Grund hierfür ist, dass die aggregierte Information den Individualdaten zugespielt wird und sich nach dem *„collapse“*- Befehl, der die Aggregation bewirkt, bereits im Arbeitsspeicher (= Masterfile) befindet. Somit muss der Individualdatensatz (*filename1*) nicht erneut geladen werden. Jedoch müssen die Beobachtungen

nach dem „merge"- Befehl erneut nach den Schlüsselvariablen sortiert werden, daher zeichnet sich die Alternativlösung durch insgesamt nur einen Befehl weniger aus.[14]

Ein Vorteil, den *„merge"* gegenüber *„egen"* hat, würde generell nur dann zu Tage treten, falls die Fälle nicht immer eindeutig einer Einheit zugeordnet werden könnten. Dies liegt daran, dass mit der Option *„nokeep"* nicht zuweisenbare Beobachtungen im Using-File gelöscht werden können oder mit der Option *„unique"* nur Beobachtungen zusammengespielt werden, die über die Schlüsselvariable(n) eindeutig identifizierbar sind. Im Mikrozensus ist eine eindeutige Zuordnung jedoch immer anhand der Variablen *EF1, EF3* und *EF4* (Konzept des Haushaltes) aufgrund ihres hierarchischen Aufbaus möglich.

3 Gegenüberstellung von STATA- und SPSS-Syntax

In folgender Tabelle sind die drei möglichen Befehlssequenzen in STATA zu *„merge"* und *„egen"* und ihre Entsprechung in SPSS anhand der Rekonstruktion der Variable *EF524* aufgeführt. Bei *„merge"* werden sowohl die Version, bei der die Individualdaten den Aggregatdaten zugespielt werden, als auch die Alternative, bei der die Aggregatdaten den Individualdaten zugepielt werden, gegenübergestellt. Insbesondere wird deutlich, welche Operationen *„egen"* in nur einem Befehl ausführen kann, worin sich die Eleganz dieses Befehls ausdrückt .[15]

[14] Es empfiehlt sich, den Unterschied der Syntax im kommentierten Anhang unter 2a) und 2b) nachzuvollziehen.

[15] Auch bei der Rekonstruktion der Variable *EF557* hat sich der *„egen"*-Befehl bewährt. Im Anhang wird dargestellt, wie man auch diese Variable mit den verschiedenen Befehlen rekonstruieren kann.

Abb. 5: Gegenüberstellung der Befehlssequenzen in STATA und SPSS

STATA			**SPSS: *„aggregate“*[16]**
Befehl *„egen“*	**Befehl *„merge“***		
	I → A[17]	**A → I**[18]	
	sort EF1 EF3 EF4	sort EF1 EF3 EF4	sort cases by EF1 EF3 EF4.
egen v524 = sum(EF504==2 & EF506==1), by(EF1 EF3 EF4)	save *filename1*, replace	save *filename1*, replace	
	gen v524 = EF504==2 & EF504<=2 & EF506==1	gen v524 = EF504==2 & EF504<=2 & EF506==1	compute x524=0. if (EF504=2 & EF506=1) x524=1.
	collapse (sum) v524, by (EF1 EF3 EF4)	collapse (sum) v524, by (EF1 EF3 EF4)	aggregate outfile=* mode=addvariables /presorted /break EF1 EF3 EF4 /v524'Zahl Ilo-Erwerbslose im Haushalt, Privathaushalt' = sum(x524).
	sort EF1 EF3 EF4	sort EF1 EF3 EF4	
	save *filename2*, replace		
	use *filename1*, clear		
	merge EF1 EF3 EF4 using *filename2*, uniqusing	merge EF1 EF3 EF4 using *filename1*, uniqusing	
		sort EF1 EF3 EF4	
tab v524 EF524 if EF506==1, miss	tab v524 EF524 if EF506==1, miss	tab v524 EF524 if EF506==1, miss	crosstabs ef524 by v524 /missing include.

[16] Diese Syntax wurde anhand der SPSS-Version 15.0 getestet.
[17] Zuspielen der **I**ndividualdaten zu den **A**ggregatdaten
[18] Zuspielen der **A**ggregatdaten zu den **I**ndividualdaten

Literatur

Kohler, U., und Kreuter, F, 2006: Datenanalyse mit STATA. . 2. Auflage. München: Oldenbourg

Kohler, U., und Kreuter, F, 2001: Datenanalyse mit STATA. 1. Auflage. München: Oldenbourg

Lengerer, A., und Boehle, M., 2006: Rekonstruktion zu Bandsatzerweiterungen zu Haushalt, Familie und Lebensformen im Mikrozensus. GESIS-ZUMA-Methodenbericht 2006/05. Mannheim.

Anhang

A1

STATA-Syntax zur Rekonstruktion der Bandsatzerweiterung *EF524*.

```
* version 9.2 or version 10
clear
capture log close
cd Laufwerk:\<Padangabe>
log using rekonstruktion_ef524.log, text replace
set more off
set mem 500m
set dp comma
use EF1 EF3 EF4 EF504 EF506 EF524 using "mz2004.dta", clear
set more off
* Sortierten Datensatz für den „merge“-Befehl und den
* „joinby“-Befehl speichern
sort EF1 EF3 EF4
save Laufwerk:\<Padangabe>\filename1.dta, replace

* 1. Replikation der Variable EF524 "Anzahl der ILO-Erwerbslosen im
* Haushalt" durch „egen“

* Einschränkung auf Personen in Privathaushalten EF506==1
egen v524 = sum(EF504==2 & EF506==1), by(EF1 EF3 EF4)
label variable v524 "Anzahl ILO-Erwerbslose im Privathaushalt"

* 2. Replikation der Variable EF524 "Anzahl der ILO-Erwerbslosen im
* Haushalt" durch „merge“

* 2a) 1. Möglichkeit: Zuspielen der Individualdaten zu den Aggregatdaten
sort EF1 EF3 EF4
gen v524 = EF504==2 & EF504<=2 & EF506==1
* Aggregieren auf Haushaltsebene
collapse (sum) v524, by(EF1 EF3 EF4)
sort EF1 EF3 EF4
save Laufwerk:\<Padangabe>\filename2, replace
* Individualdatenfile=filename1.dta als Masterfile einlesen
```

```
use Laufwerk:\<Padangabe>\filename1, clear
* Individualdaten filename1.dta als Masterfile mit Aggregatdaten "filename2.dta"
* zusammenführen
merge EF1 EF3 EF4 using Laufwerk:\<Padangabe>\filename2.dta, uniqusing
```

*** 2b) 2. Möglichkeit: Zuspielen der Aggregatdaten zu den Individualdaten**

```
sort EF1 EF3 EF4
gen v524 = EF504==2 & EF504<=2 & EF506==1
* Aggregieren auf Haushaltsebene
collapse (sum) v524, by(EF1 EF3 EF4)
* bei Version 2a), ergo dem Zuspielen der Individualdaten zu den
* Aggregatdaten müssten die folgenden zwei Befehle auch
* ausgeführt werden:
* sort EF1 EF3 EF4
* save p:\<Padangabe>\filename2.dta, replace
sort EF1 EF3 EF4
merge EF1 EF3 EF4 using Laufwerk:\<Padangabe>\filename1.dta, uniqusing
sort EF1 EF3 EF4
```

*** 3. Replikation der Variablen *EF524* mithilfe des *„joinby“*- Befehls**

```
* Der Befehl „joinby“ erzeugt eine Bandsatzerweiterung, indem er durch die
* Sortierung nach Variablen erzeugte "Gruppen" verbindet.
sort EF1 EF3 EF4
gen v524 = EF504==2 & EF504<=2 & EF506==1
* Aggregieren auf Haushaltsebene
collapse (sum) v524, by(EF1 EF3 EF4)
sort EF1 EF3 EF4
save p:\<Padangabe>\filename3.dta, replace
sort EF1 EF3 EF4
joinby EF1 EF3 EF4 using Laufwerk:\<Padangabe>\filename1.dta, unmatched(both)
sort EF1 EF3 EF4
* Um bei den drei Möglichkeiten zu überprüfen, ob die richtige Anzahl an
 * Erwerbslosen für die einzelnen Haushalte gebildet wurde, dienen der „list“-
* und „tab“- Befehl für die jeweiligen Variablen:
list EF1 EF3 EF4 EF504 EF506 EF524 v524 in 1/25, nolab sepby(EF1 EF3 EF4)
tab v524 EF524 if EF506==1, miss
```

*** Alle Befehle liefern dasselbe Ergebnis

SPSS-Syntax zur Rekonstruktion der Bandsatzerweiterung *EF524*.

```
* version 15.0.
* Bei der SPSS-Syntax muss beachtet werden, dass der „aggregate“-Befehl immer eine.
* Hilfsvariable (hier: x524) durch eine neue Variabel (hier:v524) ersetzt.

GET FILE='mz2004.sav'
    /keep EF1 EF3 EF4 EF504 EF506 EF524.
sort cases by EF1 EF3 EF4.

compute x524=0.
if (EF504=2 & EF506=1) x524=1.
* Doppelzählungen möglich, da Personen sowohl.
* am Hauptwohnsitz als auch am Nebenwohnsitz befragt werden können.
aggregate outfile=* mode=addvariables
    /presorted
    /break EF1 EF3 EF4
    /v524 'Zahl ILO-Erwerbslose im Haushalt, Privathaushalt' = sum(x524).
* Kreuztabellierung von Variable EF524 und der selbst generierten
* Bandsatzerweiterung v524.
crosstabs ef524 by v524
/missing include.
* Da sich eine Diagonale ergibt, entspricht v524 der Variablen EF524.
```

A2

STATA-Syntax zur Rekonstruktion der Bandsatzerweiterung *EF557*.

```
* version 9.2 or version 10
clear
capture log close
cd Laufwerk:\<Padangabe>
log using rekonstruktion_ef557.log, text replace
set more off
set mem 500m
set dp comma
use EF1 EF3 EF4 EF32 EF506 EF507 EF557 using "mz2004.dta", clear
set more off

* Sortierten Datensatz für den merge-Befehl und den
```

```
* joinby-Befehl speichern
sort EF1 EF3 EF4
save Laufwerk:\<Padangabe>\filename4.dta, replace
```

*** 1. Replikation der Variable *EF557* "Geschlecht: Haushaltsbezugsperson" durch „*egen*“**

```
* Erzeugen der Variablen n_bezper_hh
* „Anzahl der Bezugspersonen im Haushalt“
egen n_bezper_hh = sum(EF506==1 & EF507==1), by(EF1 EF3 EF4)
label variable n_bezper_hh „Anzahl der Bezugspersonen im Haushalt“
tab n_bezper_hh

* Falls sich in einem Haushalt mehrere Bezugspersonen befinden sollten, muss die re-
* konstruierte Bandsatzerweiterung v557„Geschlecht: Haushaltsbezugsperson“ als Missing
* deklariert werden. Dies kann man in einem Befehl erledigen.

if n_bezper_hh==1 {
          egen v557 = max(EF32 * (EF506==1 & EF507==1) ), by(EF1 EF3 EF4)
}
else {
        replace v557==.
       }
label variable v557„Geschlecht: Haushaltsbezugsperson“

* Bei der Benutzung der geschweiften Klammern „{}“ nach dem „if-“ bzw.“else-“Befehl, ist
* zu beachten, dass die öffnende Klammer „{„ in derselben Zeile wie der „if-“ bzw
* „else-„ Befehl stehen muss und das darauffolgende Kommando erst in der nächsten
* Zeile unter halb der Klammer stehen kann. Die schließende Klammer } soll ihrerseits
* in einer neuen Zeile unterhalb des letzten Kommandos stehen.
* Kreuztabellierung von Variable EF557 und der selbst generierten.
* Bandsatzerweiterung v557.
tab v557 EF557
*Da sich eine Diagonale ergibt, entsprechen die beiden Variablen einander.
*Kontrollliste
list EF1 EF3 EF4 EF32 EF506 EF507 v557 EF557 n_bezper_hh, ///
 nolab sepby(EF1 EF3 EF4)
*Analog können wiederum die durch die folgenden Befehle erzeugten Ergebnisse
*überprüft werden.
```

*** 2. Replikation der Variable *EF557* "Geschlecht: Haushaltsbezugsperson"**
durch *„merge“*

*** 2a) 1. Möglichkeit: Zuspielen der Individualdaten zu den Aggregatdaten**

```
sort EF1 EF3 EF4
* Um die Fälle, bei denen die Anzahl der Bezugspersonen in einem Haushalt größer
* als 1 ist, als Missing zu deklarieren, müsste die Bandsatzerweiterung "Anzahl der
* Bezugspersonen in einem Haushalt" zunächst einmal generiert werden.
*Wird der „egen-" Befehl nicht verwendet, kann man diese Variable mithilfe der
* "sum“-Funktion beim "collapse“-Befehl analog wie bei der beschriebenen
* Rekonstruktion der Bandsatzerweiterung "Erwerbslose
*(EU-Definition) im Haushalt: Anzahl" EF524 generieren.

gen v557 = EF32 * (EF506==1 & EF507==1)
* Aggregieren auf Haushaltsebene
collapse (max) v557, by(EF1 EF3 EF4)
* benutzt man beim „collapse-„ Befehl statt der Maximierungsfunktion die Summierungsfunktion:
* collapse (sum) v557, by(EF1 EF3 EF4), erhält man dasselbe Ergebnis
sort EF1 EF3 EF4
save Laufwerk:\<Padangabe>\filename5.dta, replace
* Individualdatenfile=filename4.dta als Masterfile einlesen
use Laufwerk:\<Padangabe>\filename4.dta, clear
* Individualdaten filename4.dta als Masterfile mit Aggregatdaten "filename5.dta"
* zusammenführen
merge EF1 EF3 EF4 using Laufwerk:\<Padangabe>\filename 5.dta, uniqusing
tab v557 EF557
*Kontrollliste
list EF1 EF3 EF4 EF32 EF506 EF507 v557 EF557, ///
nolab sepby(EF1 EF3 EF4)
```

*** 2b) 2. Möglichkeit: Zuspielen der Aggregatdaten zu den Individualdaten**

```
sort EF1 EF3 EF4
gen v557 = EF32 * (EF506==1 & EF507==1)
*Aggregieren auf Haushaltsebene
collapse (max) v557, by(EF1 EF3 EF4)
sort EF1 EF3 EF4
```

```
merge EF1 EF3 EF4 using Laufwerk:\<Padangabe>\filename4.dta
sort EF1 EF3 EF4
```

*** 3. Replikation der Variablen *EF524* mithilfe des „*joinby*“-Befehls**

```
sort EF1 EF3 EF4
gen v557 = EF32 * (EF506==1 & EF507==1)
*Aggregieren auf Haushaltsebene
collapse (max) v557, by(EF1 EF3 EF4)
sort EF1 EF3 EF4
save Laufwerk:\<Padangabe>\filename6.dta, replace
sort EF1 EF3 EF4
joinby EF1 EF3 EF4 using Laufwerk:\<Padangabe>\filename4.dta, unmatched(both)
sort EF1 EF3 EF4
```

SPSS-Syntax zur Rekonstruktion der Bandsatzerweiterung *EF557*[19]

```
GET FILE='mz2004.sav'
 /keep EF1 EF3 EF4 EF32 EF506 EF507 EF557.
sort cases by EF1 EF3 EF4.
* Auch bei der SPSS-Syntax muss darauf geachtet werden, dass nur eine Person im Haushalt die.
* Bezugsperson ist.
compute n_bezugspersonen_hh=0.
if (EF506=1 & EF507=1) n_bezugspersonen_hh=1.
aggregate outfile=* mode=addvariables
    /presorted
    /break EF1 EF3 EF4
    /n_bez_hh 'Zahl der Bezugspersonen im Haushalt' = sum(n_bezugspersonen_hh).
keep if n_bez_hh ==1.
compute x557=0.
if (EF506=1 & EF507=1) x557=EF32.
sort cases by EF1 EF3 EF4.
aggregate outfile=* mode=addvariables
 /presorted
 /break EF1 EF3 EF4
```

[19] Neben der hier dargestellten „aggregate“-Variante besteht in SPSS zudem die Möglichkeit, ein outfile zu erzeugen und dann zu matchen. Vgl. Lengerer und Boehle 2006: 10.

```
 /v557 'Geschlecht: Haushaltsbezugsperson' = sum(x557).
*Kreuztabellierung von Variable EF557 und der selbst- .
*generierten Bandsatzerweiterung v557.
crosstabs EF557 by v557
 /missing include.
```